View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

Hrvatska revija za rehabilitacijska istraživanja 2017, Vol 53, br. 2, str. 61-71

CroDA: A CROATIAN DISCOURSE CORPUS OF SPEAKERS WITH APHASIA

JELENA KUVAČ KRALJEVIĆ[1], GORDANA HRŽICA[1], KAROLINA LICE[2]

[1]Department of Speech and Language Pathology, Faculty of Education and Rehabilitation Sciences, University of Zagreb, [2]SUVAG Polyclinic, Zagreb. Contact: jkuvac@erf.hr

Received: 11.05.2017.
Accepted: 26.09.2017.

Original scientific paper
UDK: 811.163.42 '234.2

Abstract: *The paper describes data collection and transcription to develop the Croatian discourse corpus of speakers with aphasia (CroDA), developed within the framework of the project Adult Language Processing (HRZZ-2421-UIP-11-2013) and available from 2017 as part of the AphasiaBank database of multimedia interactions for studying communication among speakers with aphasia. In accordance with the AphasiaBank Protocol, the following discourse tasks were sampled: personal narrative, picture description, story narrative and procedural discourse. Recorded speech was transcribed according to the Codes for Human Analysis of Transcripts (CHAT). CroDA, as the first discourse corpus of speakers with aphasia in Croatian, may provide new insights into specific linguistic features of discourse produced by speakers with aphasia and serve as a useful resource for quantitative and qualitative analysis.*

Keywords: *aphasia, CroDA, AphasiaTalk, Croatian, discourse*

INTRODUCTION

Corpora are regarded as objective and comprehensive linguistic resources for analysing written or spoken language in formal or informal situations. However, few corpora exist of language spoken by those with developmental or acquired language disorders such as aphasia. Aphasia is an acquired neurological disorder caused by stroke in 80% of cases, and in other cases by tumors, injuries and infections. Among patients with first-ever ischemic stroke (FEIS), aphasia occurs in 15% of those younger than 65 and in 43% of those 85 or older (Engelter et al., 2006). This high prevalence has made aphasia an increasingly important topic of neurolinguistic and psycholinguistic research.

Aphasia symptomatology varies with lesion location and size, giving rise to several subtypes of aphasia, which can be classified as "fluent" or "nonfluent" (Spreen and Risser, 2003). In fluent aphasia such as Wernicke`s aphasia, spoken utterances are relatively fluent but are difficult to understand because they contain invented or irrelevant words. Nonfluent aphasia, such as Broca's aphasia, features significantly impaired production of spoken language but relatively good comprehension (see Hedge, 2006). Goodglass and Kaplan (1983) described fluent aphasic speech as speech delivered at a normal or excessive speed and featuring normal phrase length, melody, rhythm, articulatory agility and normal or paragrammatic form. They described nonfluent aphasic speech as slow speech featuring reduced phrase length, abnormal intonational contour, effortful articulation and simplified syntax and/or absent grammar. Despite the existence of many classifications of aphasia, the distinction between "fluent" and "nonfluent" speech is of critical value when characterising speech production and is often used as an effective "first cut" in diagnostic classification (Kerschensteiner et al., 1972, Goodglass et al., 1964, Benson, 1967).

Although language of speakers with aphasia have been studied linguistically since the 19th century, corpus-based studies of such speakers are relatively recent; such studies aim to directly address "the question of how aphasic speakers talk in connected speech in everyday situations" (Armstrong 2000, p. 875). Corpora are indispensable for analysis of discourse, defined by functionalist linguistics as *language-in-use* (Halliday and Hassan, 1976,

Schriffin, 1994), or by formalist linguistics as *the unit above a sentence* (see Harris, 1976). For formalists, discourse analysis allows various microstructural linguistic analyses; for functionalists, it allows macrostructural analyses of discourse structure, topic maintenance and turn-taking. To describe the discourse ability of speakers with aphasia comprehensively, researchers can combine the functionalist and formalist approaches in a so-called "multi-level" approach that examines linguistic and cognitive functions (see Marini et al., 2011; Linnik et al., 2016).

Corpora of speakers of aphasia

Corpora of speakers with aphasia reflect a long, fruitful tradition of speech sampling, transcription, and annotation with available coding systems. However, design and annotation protocols in corpora of speakers with typical language status cannot be applied directly to corpora of speakers with aphasia given the latter's special character, but they can be used as a starting point. Established corpora start by selecting existing transcription and annotation procedures and building upon them in order to capture relevant characteristics of speech of speakers with aphasia. These procedures are regularly described and published as a part of the corpora or as a separate paper.

Specialised corpora of speakers with aphasia have been established in some languages, including Dutch (CoDAS; Westerhout, Monachesi, 2006), Greek (GREECAD, Varlokosta et al., 2016), and Russian (CliPS - Clinical Pear Stories; Khudyakova et al., 2016). CliPS and GREECAD are being annotated using ELAN software (https://tla.mpi.nl/tools/tla-tools/elan/), developed by the Max Planck Institute for Psycholinguistics and The Language Archive in Nijmegen, The Netherlands (Brugman and Russel, 2004). This software allows transcription in separate tiers, thus enabling annotated speech samples, with the options of encoding information at the microlinguistic level (e.g. part of speech tagging, grammatical errors, clause types) and discourse level (e.g. narrative structure elements, main events, evaluation devices). Additional information about the annotation procedure is available within these corpora. CoDAS relies on existing Dutch spoken language corpora and child language corpora transcribed using annotation tools developed as part of the CHILDES project (MacWhinney, 2000). The CHILDES database, with its coding system Codes for Human Analysis of Transcripts (CHAT) and annotation and analysis program Computerised Language Analysis (CLAN), plays a central role in child language corpora research. Subsequently, it was expanded to accommodate a broader range of corpora, and today CHILDES is a part of the TalkBank (MacWhinney, 2007), which contains diverse language sample banks, including AphasiaBank (MacWhinney at al., 2011), the largest database for communication studies on speakers with aphasia. Corpora in the following languages are available as part of AphasiaBank: English, German, French, Cantonese, Italian, Japanese, and Spanish.

Most of the corpora of speakers with aphasia focus on discourse production by patients with aphasia caused by left-hemisphere damage. Although individuals with right-hemisphere damage usually do not show language problems at the microlinguistic level, they show some difficulties in language comprehension and production at the discourse level (Tompkins et al., 1997, Khudyakova et al., 2016). At least one corpus, CliPS, includes language samples of speakers with right-hemisphere damage (Khudyakova et al., 2016).

AphasiaBank

TalkBank is a large database of spoken-language corpora covering different languages (https://talkbank.org, MacWhinney, 2007); it contains corpora from several subfields of communication, including first- and second- language acquisition, conversation analysis and clinical communication. AphasiaBank (http://aphasia.talkbank.org/) is a subsection of TalkBank that contains multimedia interactions for communication studies in aphasia. Established by Brian MacWhinney and Audrey Holland in 2007, it "has been designed to replicate and extend the organisational model established by the Child Language Data Exchange System (CHILDES) for the field of child language acquisition" (MacWhinney et al., 2011, p. 1287). AphasiaBank is based on structured interviews of speakers with aphasia conducted by clinicians (MacWhinney et al., 2011). The use of structured

interviews is meant to ensure consistent experimental conditions and comparability across participants and languages. AphasiaBank contains narrative, procedural, personal, and descriptive discourse from 290 persons with aphasia, as well as 190 control participants (MacWhinney and Fromm, 2016). Participants are primarily individuals whose aphasia resulted from a stroke that was verified through neuroimaging or definitive medical diagnosis. Participants with dementia or with comorbidities seriously affecting cognitive function were excluded.

Most corpora in AphasiaBank were collected using a standardised data collection protocol, maximising comparability across participants and languages. The protocol consists of four discourse elicitation tasks: personal narratives, picture descriptions, storytelling, and procedural discourse.

AphasiaBank grows as new data from new clinical groups and languages are contributed, as reflected in the recent addition of TBIBank, a longitudinal database of speakers with aphasia video-recorded at six different time points following traumatic brain injury in order to track recovery; and of DementiaBank, which assesses the ability of people with dementia to perform a uniform set of tasks (Forbes et al., 2012). The research and clinical applications of AphasiaBank also grow as new algorithms and linguistic parameters are developed. For example, Leonid Spektor developed the EVAL program in 2012 to serve speech language pathologists. EVAL produces a language profile from transcriptions created in the CHAT format. The language profile includes numerous assessment measures: number of utterances, mean length of utterances (MLU) in words and in morphemes, type/token ratio (TTR), average number of clauses per utterances, number of errors on word and utterance level, number of repetitions, number of self-corrections, duration of speech, parts of speech (nouns, verbs, adjectives, adverbs), verb tenses and plurals (Forbes et al., 2012; Talkbank manual). EVAL generates the language profile as an Excel spreadsheet. These profiles can be compared between a given speaker with aphasia and control individuals from AphasiaBank, or between two or more transcripts of the same speaker with aphasia.

CroDA

Lack of resources is an obstacle in the study of speech of speakers with aphasia in Croatian, which requires diverse types of sampling. Discourse samples of speakers with aphasia enable research in all language domains, and audio and video files enable research on the characteristics of aphasic speech. Collecting these types of data increases comparability of the data and allows creation of open access corpora to benefit the wider research community.

The goal of this paper is to describe the preparation, participants, data collection, transcription and annotation of the Croatian discourse corpus of speakers with aphasia (CroDA), which launched on AphasiaBank in 2017. The paper also aims to provide a set of specifications that would serve for the development of other relevant corpora. Such resources would contribute to the future research of language of speakers of aphasia in Croatia, and it could provide guidelines for corpus design in other languages.

AphasiaBank tools and protocols were selected as the starting point for CroDA formation because: a) the elicitation procedure of AphasiaBank enables cross-linguistic research; b) AphasiaBank provides open access to researchers, and it enjoys an established presence in the international community; and c) two Croatian corpora in Croatian are already available within TalkBank, i.e. the Croatian Corpus of Child Language (Kovačević, 2002) and the Croatian Adult Spoken Language Corpus (HrAL; Kuvač Kraljević and Hržica, 2016), which might serve as a starting point regarding methodological procedures. An additional argument for selecting CLAN as a transcription and coding tool is its compatibility with other relevant programs such as Praat (Boersma, 2001; Boersma and Weenink, 2017) and ELAN (more information can be found at http://talkbank.org/software/).

Special emphasis will be given to upgrades in transcription and annotation in CroDA, especially regarding a) segmentation, b) coding, c) error marking and d) coding fluency in speech. Emphasis will also be given to language sampling differences from procedures used in previous Croatian spoken corpora.

This corpus will supplement the corpora in other languages to serve researchers, speech language

Table 1. *Demographic data on CroDA participants.*

Aphasia type	Participant	Gender	Age	Years of education	Time since brain damage onset	Therapy duration
Fluent aphasia	ZH	M	76	16	0;6	0;5
	DP	F	58	16	2;11	1;1
	MH	F	77	11	1;4	0;9
	DP	M	52	10	1;8	1;0
	PZ	M	79	11	0;10	0;8
	SV	F	59	12	5;11	5;1
	MH	M	58	12	5;10	5;0
	MĆ	M	55	16	0;8	0;2
	IS	M	74	16	0;10	0,2
	NM	M	64	16	0;6	0;2
Non-fluent aphasia	VP	M	64	12	4;3	3;7
	HK	F	77	11	0;9	0;2
	DT	M	52	12	0;4	0;2
	DD	F	72	12	0;10	0;6
	DL	M	48	11	1;9	1;5
	NB	M	62	11	6;5	3;0
	LKH	F	79	16	1;6	0;3
	MB	M	65	16	0;6	0;2
	BD	F	72	8	1;0	0,9
	IB	M	50	16	0;8	0;2

pathologists and others in exploring language and communication skills of people affected by aphasia (MacWhinney, Fromm, 2016). Like other corpora, CroDA can be downloaded and transcripts can be browsed following user registration. CroDA, developed within the project *Adult Language Processing (HRZZ-2421-UIP-11-2013)*, is the first specialised corpus that contains discourse samples of Croatian speakers with aphasia.

PARTICIPANTS

For this study, 20 patients with aphasia, diagnosed at the Polyclinic for the Rehabilitation of Listening and Speech (SUVAG) by a speech-language pathologist, were recruited. They were all monolingual and right-handed prior to the event that caused their aphasia. All patients developed classical aphasia symptoms as a result of stroke (18 participants), brain injury (1 participant) or tumor (1 participant). Patients were additionally classified as fluent or nonfluent according to the main characteristics of speech fluency, as proposed by Goodglass and Kaplan (1983). All participants signed an informed consent form according to the Helsinki Ethical Principles for Medical Research. Information for consent was delivered in writing and orally, with the assistance of a speech-language pathologist and family members when needed. Study procedures were approved by the Ethics Committee of the Faculty of Education and Rehabilitation Sciences. As of May 2017, CroDA includes discourse samples from 10 speakers with fluent aphasia and 10 with non-fluent aphasia, all of whom suffered left-hemisphere brain damage (Table 1).

DATA COLLECTION

In accord with the AphasiaBank protocol[1], discourse was sampled along the following four genres, which capture "different ways of using language to achieve different culturally established tasks" (Eggins and Martin, 1997, p. 9):

1. Personal narratives were obtained by asking participants to recount something about their speech, stroke and recovery (stoke story) and

1 Stimuli for elicitation of speech samples, as well as relevant instructions, can be retrieved at http://aphasia.talkbank.org/protocol/ [Accessed 12 September 2017].

to describe one important event in their lives (positive or negative).
2. Picture descriptions were elicited by showing participants three stories in the form of black-and-white drawings: "Broken window" (four panels), "Refused Umbrella" (six panels) and "Cat Rescue" (one panel). Participants were asked to look at the panels of each story and to tell a story with a beginning, middle and end.
3. Storytelling was elicited by showing participants the book *Cinderella,* with the text covered up. Participants were then asked to recount the story without looking at the book.
4. Procedural discourse was elicited by asking participants to describe how to prepare a ham-and-cheese sandwich.

A variety of genres are included to provide a well-rounded assessment of speakers' abilities to understand and produce language in various everyday settings. Selection of elicitation tasks is a highly relevant methodological question in the research of aphasia. There are two reasons for that status. First, different genres have proven to impose different linguistic and cognitive demands (Linnik et al., 2016). For example, narrative samples obtained by using a picture sequence have been shown to be richer than those elicited by a single picture (Marini et al., 2005; Olness, 2006), and there seems to be a difference in performance between personal narratives and picture-elicited narratives (Wright and Capilouto 2012a, 2012b). In addition, previous research of discourse of speakers with aphasia has been conducted using a variety of elicitation tasks: picture description, personal narratives and procedural discourse (overview: Linnik et al., 2016). Usage of different tasks limits the comparability and generalisability of findings between studies as well as between these populations of speakers and the general population. Tasks in the protocol of AphasiaBank have been selected in order to maximise the amount of information collected from participants and to allow for further research on the impact of genre on linguistic performance. The selected types of tasks have been successfully used in this population and enable comparison with previous research and other corpora.

The AphasiaBank protocol allows investigators to ask simpler questions to encourage participants to speak in the event that they fail to respond to the initial prompt within 10 seconds, and these additional prompts were required for some CroDA participants. Participants who did not produce utterances even after these additional prompts were excluded from the sampling.

Four investigators participated in the data collection. All investigators were presented with the protocol documentation (including interview protocol and troubleshooting section) translated and/or adapted to Croatian. The protocol was prepared by the authors of the corpora, and this work involved additional decisions about sample collection. Each investigator received individualised training guided by at least one of the authors, in which he or she was presented with a short overview of the CroDA, AphasiaBank, the protocol and troubleshooting.

Due to the severity of difficulties of speakers, investigators were instructed to record the protocol in several sessions if needed, in contrast to the recommendations of the AphasiaBank protocol. This resulted in several (up to four) sessions per participant.

TRANSCRIPTION AND ANNOTATION

All data were recorded in audio and video formats to ensure accurate transcription, but only audio data are available for sharing on AphasiaBank. Language samples were transcribed and coded according to the TalkBank standard rules of contribution, using the TalkBank toolkit for transcription, editing and analysis, including the CHAT transcription format and CLAN suite of programmes (MacWhinney, 2007). Transcribing spontaneous speech is challenging because it is not as fluent as written language, and it features pauses, false starts and vague utterance borders; transcribing spontaneous speech by speakers with aphasia is even more challenging (see Westerhout and Monachesi, 2006). CHAT codes capture a range of conversational structures as well as morphosyntactic and discourse features (e.g. repetitions, pauses, and errors).

CLAN allows for transcription on different levels by using dependent tiers, including phonological, morphological, and syntactic tiers, as well as error marking and gestures. For CroDA, only the

main tier, general comment tiers (situation, comment) and errors were used.

Example 1. *Extract from a coded transcript (INV – investigator, ZVO – target speaker).*

Transcript	Translation
*INV: kakav vam je govor ovih dana?	*INV: how is your speech these days?
*ZVO: &-e govor ovih dana mi je:.	*ZVO:&-e these days my speech is.
*ZVO: (.) &-a &-o &-a xxx smijem reći(i) li-la.	*ZVO:(.) &-a &-o &-a xxx I can say so-so.
*ZVO: nekad je bolji.	*ZVO:sometimes it's better.
*ZVO: nekad je: <užasan loši> [*].	*ZVO:sometimes it's <terribly bad> [*]
%err: užasan loši = užasno loš $MOR $NFL	

Prior to the transcription, 35 CHAT conversational codes that have been used most extensively during the transcription of the Croatian Adult Spoken Language Corpus (Kuvač Kraljević and Hržica, 2016) were selected. During the transcription, seven new codes were added to the list (marked with an asterisk in Table 2). These codes were relevant for the CroDA and are presented in Table 2.

Error coding consisted of three procedures. First, transcribers marked all words, phrases or sentences that they found to be erroneous. Second, authors examined marked elements and decided whether the decision was correct (excluding, for example, dialectal forms). Third, errors were coded according to CHAT. First-level code marks the language level (e.g. morphological error), and second-level code marks the type of error. Error codes are presented in Table 3.

Speech samples were transcribed by a group of 20 transcribers supervised by the authors of the corpus. Transcribers received approximately 8 hours of training in transcription and coding prior to transcription. They also received printed transcription instructions and were continuously supervised during transcription and coding. Speech streams were segmented into communication units (C-units; Loban, 1966) based on syntactic criteria. The C-unit is based on the T-unit, which is defined as the "one main clause plus whatever subordinate clauses happen to be attached or embedded within it" (Hunt, 1966, p. 735). Thus, C-units include T-units as well as isolated phrases unaccompanied by a verb. Such phrases typically appear in answers to questions (Crookes, 1990). Supervision included two individual feedback sessions for each transcriber, six hours of group sessions in which different issues were addressed and an online platform (message board) that allowed additional discussions. Each transcript was re-checked twice by experienced researchers, once for C-unit segmentation and again for coding. Each transcript was also verified using the CHECK programme of the CLAN suite. This process allowed for high inter-annotator reliability. Approximately 10% of each speech sample was extracted, re-transcribed and re-coded by another transcriber or one of the authors. The two transcriptions showed an average match of 89% (minimum 82%, maximum 100%).

Table 3. *Codes used to mark speakers' errors.*

First-level code	Description	Second-level code	Description
$PHO	phonological errors	$ADD	addition
		$LOS	loss
		$SUB	substitution
$LEX	inappropriate choice of word		
$SYN	syntactic errors		includes errors in word order, clause order or discourse syntax
$MOR	morphological errors	$PRE	error in the derivation of prefixed word
		$SUF	error in the derivation of suffixed word
		$NFX	error in the derivational infix
		$DER	other derivational errors (e.g. in compounding)
		$REG	overgeneralisations
		$NFL	inflectional errors

Table 2. *Codes used for the transcription of CroDA*

Type of event	Event	Main line	Dependent tier
Speech events	lengthening of a consonant or vocal	:	-
	*pause	(.) (..) (...)	
	filled gaps	&-'phonetic transcription'	
	phonological fragments	&+'phonetic transcription'	
	*broken word	^	
	oscillations in speech volume	[=! yelling] [=! whispering]	(or) described in Comment tier
Paralinguistic material	laughter	[=! laughing]	(or) described in Comment tier
	cough	[=! coughing]	(or) described in Comment tier
Repetitions and reformulations	whole word repetition	[/]	-
	phrase repetition	(phrase embedded in <>)+[/]	
	word revision	[//]	
	phrase revision/retracing	(phrase embedded in <>)+[//]	
	reformulation	[///]	
Interruptions and overlaps	trailing off	+...	-
	interruption	+/. +/? +/!	
	*self-interruption	+//. +//? +//!	
	overlap	+<	
Continuation after interruption	self completion	+,	-
	completion of other speaker	++	
Quotation	quotation follows	+"/.	-
	quotation precedes	+".	
Incomplete and unintelligible words	unintelligible word	xxx	-
	*best guess	[?]	
	omitted word	0word	
	omitted auxiliary	0aux	
	*omitted preposition	0prep	
	omitted verb	0v	
	*omitted reflexive pronoun	0ref	
	incomplete word/shortened word	wo(rd)	
	replacement	[: word] (replaced preceding word)	
Special form markers	dialect form	@d	-
	*neologism	@n	
	interjection	@i	
	letter	@l	
	onomatopoeia	@o	
	second-language form	@s:en (English word)	
	word play	@wp	
Comments		-	described in Comment tier
Error coding	error	[*]	(and) described in Error coding tier

Transcripts were linked to the source audio file using the Transcriber function in the CLAN suite. This allows for both continuous playback and utterance-level playback. Media files are available in the downloadable version of CroDA.

CORPUS DESCRIPTION

As of August 2017, CroDA consists of transcripts of 20 speakers with aphasia. Each speaker produced four texts elicited by four tasks. Some of the speakers were not able to complete this in only one session, so they were recorded in several sessions on different dates. Some speakers performed more than one task during one session, but the recording was interrupted. Overall, this resulted in more than one transcript per participant (up to four - see Table 4).

Table 4. *Number of transcripts per session per participant*

Aphasia type	Participant	Gender	Number of transcript
Fluent aphasia	ZH	M	2
	DP	F	3
	MH	F	2
	DP	M	3
	PZ	M	3
	SV	F	2
	MH	M	4
	MĆ	M	1
	IS	M	2
	NM	M	4
Non-fluent aphasia	VP	M	1
	HK	F	3
	DT	M	1
	DD	F	3
	DL	M	2
	NB	M	4
	LKH	F	3
	MB	M	3
	BD	F	3
	IB	M	2

Relevant demographic data are available for each participant in a separate spreadsheet that is a part of the CroDA, including age, gender and education. Relevant medical information, including type of aphasia, speech and language therapy, and time of brain damage are also part of a spreadsheet.

The corpus consists of 35,315 tokens in 7,752 communication units, corresponding to approximately 388 utterances per speaker (minimum 126, maximum 1,687) and 1,766 tokens per speaker (minimum 373, maximum 5,278).

DATA SHARING

AphasiaBank, operating under TalkBank data sharing rules, provides eight levels of data sharing, ranging from full web access with no attempt at anonymisation to archiving-only functionality (http://talkbank.org/share/irb/options.html). To protect participants in CroDA, which is especially important given the small communities where many samples were recorded, the last names and detailed addresses of participants have been withheld, and participants' first names and/or initials have been pseudonymised. Each participant is identified using a unique three-letter sequence.

The authors of CroDA are happy to share the corpus and related data with the wider community according to the basic rules of TalkBank (https://talkbank.org/share/rules.html). In particular, any published works derived from AphasiaBank corpora should cite the references listed in the manuals, as well as the present article in the case of CroDA. Any such works should also cite the original TalkBank publication (MacWhinney, 2007). CroDA is available under a *Creative Commons Attribution-ShareAlike 4.0 International Public License* (https://creativecommons.org/licenses/by-sa/4.0/legalcode).

CONCLUSION

Corpus-based research, well established in linguistics for studying language variation (Svartvik, 1992), has become one of the most widely used tools for analysing speech language pathology (Ferguson et al., 2009) in the context of naturalistic, multifunctional discourse, which can help to guide clinical practice (see Armstrong, 2000, Linnik et al., 2016).

The main rationale for developing CroDA is to facilitate studies of the overall communication abilities as well as specific linguistic features of various types of discourse produced by Croatian speakers with aphasia. CroDA is well suited for

assessing what Holland (1982) calls "functional communication", i.e. the ability of speakers with aphasia to produce discourse and communicate despite microstructural language difficulties. CroDA provides information about spoken grammar and lexicon, discourse skills, error production and productivity. It also provides information about non-linguistic abilities such as maintaining a topic, taking turns, and being informative.

Like other corpora of speakers with aphasia, CroDA is well suited for a broad range of quantitative language analyses. For example, preliminary analysis of CroDA shows that among all the discourse genres analysed, personal narrative is the genre in which speakers with aphasia are the most productive and display the greatest syntactic complexity. It is also the genre in which their lexicon is least diverse (Kuvač Kraljević, Matić, 2017). In addition to quantitative analyses, CroDA will facilitate qualitative studies examining, for example, articulation errors, syntactic errors, neologisms, and paraphasias. By combining quantitative and qualitative data, researchers can gain better insights into the underlying mechanisms of language production in aphasia. Practitioners, for their part, can exploit these insights to make their therapy sessions more effective.

The availability of the CroDA platform and Croatian-language discourse prompts will help researchers design their own studies and upload their results into CroDA. In fact, the next step in CroDA development will be to sample typical speakers matched by age, sex and education level to speakers with aphasia, which will allow normalisation of language measure for EVAL-based comparisons in Croatian. To the extent possible, CroDA should also include speakers with right-hemisphere damage in order to allow more comprehensive understanding of language processing in various neurological conditions.

CroDA will expand the tools available to researchers and practitioners for studying the communication of speakers with aphasia.

ACKNOWLEDGMENTS

CroDA was created within the project *Adult Language Processing* (*HRZZ-2421-UIP-11-2013*) funded by the Croatian Science Foundation. We are grateful to Marina Olujić for aligning CroDA participant selection and preparation of discourse prompts with AphasiaBank procedures. We thank Vlasta Vasilj, Kristina Škarić and Martina Vuković Ogrizek for data collection.

REFERENCES

Armstrong, E. (2000): Aphasic discourse analysis: the story so far. *Aphasiology, 14(9),* 875-892.

Benson, D. F. (1967): Fluency in aphasia: correlation with radioactive scan localization. *Cortex, 3(4),* 373-394.

Boersma, P., Weenink, D. (2017): Praat: doing phonetics by computer [Computer program]. Version 6.0.31, retrieved 21 April 2017 from http://www.praat.org/

Boersma, P. (2001): Praat, a system for doing phonetics by computer. *Glot International 5(9/10)*, 341-345.

Brugman, H., Russel, A. (2004): Annotating Multimedia/ Multi-modal resources with ELAN. Proceedings of LREC 2004, Fourth International Conference on Language Resources and Evaluation. 2065-2068.

Crookes, G. (1990): The utterance and other basic units for second language discourse analysis. *Applied Linguistics, 11,* 183-199.

Eggins, S., Martin, J. R. (1997): Genres and registers of discourse. In: van Dijk, T. A. (ed.), *Discourse as Structure and Process.* London: Sage. 230-256.

Engelter, S. T., Gostynski, M., Papa, S., Frei M, Born, C., Ajdacic Gross, V., Gutzwiller, F., Lyrer, P. A. (2006): Epidemiology of aphasia attributable to first ischemic stroke: Incidence, severity, fluency, etiology, and thrombolysis. *Stroke, 37,* 1379-1384.

Ferguson, A., Craig, H., Spencer, E. (2009): Exploring the Potential for Corpus-Based Research in Speech-Language Pathology. Proceedings of the HCSNet Workshop on Designing the Australian National Corpus. Somerville, MA: Cascadilla Proceedings. 30-36.

Forbes, M., Fromm, M., MacWhinney, B. (2012): AphasiaBank: A resource for clinicians. *Seminars in Speech and Language, 33,* 217-222.

Goodglass, H., Quadfasel, F. A., Timberlake, W. H. (1964): Phrase length and the type and severity of aphasia. *Cortex, 1(2),* 133-153.

Goodglass, H., Kaplan, E. (1983): *The assessment of aphasia and related disorders*. Philadelphia: Lea & Febiger. (2nd edition)

Halliday, M. A. K., Hasan, R. (1976/1993): *Cohesion in English.* London and NY: Longman.

Harris, Z. S. (1976) *Notes du cours de syntaxe.* Paris: Seuil.

Hedge, M. N. (2006): *A Course on Aphasia and Other Neurogenic Language Disorders.* NY: Thomson Delam Learning.

Holland, A. L. (1982): Observing functional communication of aphasic adults. *Journal of Speech and Hearing Disorders, 47,* 50-56.

Hunt, K. W. (1966): Recent measures in syntactic development. *Elementary English, 43,* 732-739.

Kerschensteiner, M., Poeck, K., Brunner, E. (1972): The fluency-nonfluency dimension in the classification of aphasic speech. *Cortex, 8(2),* 233-247.

Khudyakova, M., Bergelson, M., Akinina, Y., Iskra, E., Toldova, S., Dragoy, O. (2016): Russian CliPS: a Corpus of Narratives by Brain-Damaged Individuals. Proceedings of LREC 2016 Workshop. Resources and Processing of Linguistic and Extra-Linguistic Data from People with Various forms of Cognitive/Psychiatric Impairments.

Kovačević, M. (2002): Croatian child language corpus, CHILDES.http://childes.psy.cmu.edu/data/Slavic, (retrieved 4.2.2017.)

Kuvač Kraljević, J., Hržica, G. (2016): Croatian Adult Spoken Language Corpus (HrAL). *Fluminensia: Journal for philological research. 28(2),* 87-102.

Kuvač Kraljević, J., Matić, A. (2017): Croatian Discourse Corpus of Speakers with Aphasia (CroDA): overview and first data. Paper presented at 9th International Conference of the Faculty of Education and Rehabilitation Sciences University of Zagreb. 17th May 2017.

Linnik, A., Bastiaanse, R., Höhle, B. (2016): Discourse production in aphasia: a current review of theoretical and methodological challenges. *Aphasiology, 30(7),* 765-800.

Loban, W. (1966): *Language Ability: Grades Seven, Eight, and Nine.* Washington, DC: Government Printing Office.

MacWhinney, B., Fromm, D. (2016): AphasiaBank as BigData. *Seminars in Speech and Language, 37(1),* 10-22.

MacWhinney, B., Fromm, D., Forbes, M., Holland, A., (2011): AphasiaBank: Methods for studying discourse. *Aphasiology, 25(11),* 1286-1307.

MacWhinney, B. (2007): The TalkBank Project. In: J. C. Beal, K. P. Corrigan, H. L. Moisl (eds.), *Creating and Digitizing Language Corpora: Synchronic Databases,* Volume 1. Houndmills: Palgrave-Macmillan. 163-180.

MacWhinney, B. (2000): *The CHILDES Project: Tools for analyzing talk.* Mahwah, NJ: Lawrence Erlbaum Associates (3rd edition).

Marini, A., Andreetta, S., de Tin, S., Carlomagno, S. (2011): A multi-level approach to the analysis of the narrative language in aphasia. *Aphasiology, 25(11),* 1372-1392.

Marini, A., Boewe, A., Caltagirone, C., Carlomagno, S. (2005): Age-related differences in the production of textual descriptions. *Journal of Psycholinguistic Research, 34,* 439–463.

Olness, G. S. (2006): Genre, verb, and coherence in picture-elicited discourse of adults with aphasia. *Aphasiology, 20,* 175–187.

Schriffin, D. (1994): *Approaches to Discourse.* Blackwell. Oxford.

Spreen, O. Risser, A. H. (2003): *Assessment of aphasia*. Oxford: University Press.

Svartvik J. (1992): Corpus linguistics comes of age. In: J. Svartvik, (ed.) *Directions in Corpus Linguistics.* Studies and Monographs 65. Berlin: Mouton de Gruyer, 7-13.

Tompkins, C. A., Baumgaertner, A., Lehman, M. T., Fossett, T. R. D. (1997): Suppression and discourse comprehension in right brain-damaged adults: A preliminary report. *Aphasiology, 11(4-5),* 505-519.

Varlokosta, S., Stamouli, S., Karasimos, A., Markopoulos , G., Kakavoulia, M., Nerantzini, M., Pantoula, A., Fyndanis, V., Economou , A., Protopapas, A. (2016): A Greek Corpus of Aphasic Discourse: Collection, Transcription, and Annotation Specifications. Proceedings of LREC 2016 Workshop. Resources and Processing of Linguistic and Extra-Linguistic Data from People with Various forms of Cognitive/Psychiatric Impairments. 14-21.

Westerhout, E. and Monachesi, P. (2006): A pilot study for a Corpus of Dutch Aphasic Speech (CoDAS). Proceedings of the 15 International Conference on Language Resources and Evaluation (LREC). 1648-1653.

Wright, H., Capilouto, G. J. (2012): Considering a multi-level approach to understanding maintenance of global coherence in adults with aphasia. *Aphasiology, 26,* 656–672.

CroDA: HRVATSKI DISKURSNI KORPUS GOVORNIKA S AFAZIJOM

Sažetak: *Rad opisuje postupak prikupljanja podataka i transkripciju upotrijebljenu u razvoju Hrvatskog diskursnog korpusa govornika s afazijom razvijenog u sklopu projekta Adult Language Processing (HRZZ-2421-UIP-11-2013) i dostupnog od 2017. kao dio AphasiaBank - baze multimedijalne interakcije za proučavanje komunikacije među govornicima s afazijom. U skladu s protokolom AphasiaBanka uzorkovani su diskursi na temelju četiriju zadataka: pripovijedanja osobne priče, opisa slike, prepričavanja priče i proceduralnog diskursa. Snimljeni govorni uzorci transkribirani su u skladu s Codes for Human Analysis of Transcripts (CHAT). CroDA, kao prvi diskursni korpus govornika s afazijom u hrvatskom može dati nove uvide u specifična jezična obilježja diskursne proizvodnje govornika s afazijom i poslužiti kao korisni izvor za kvantitativne i kvalitativne analize.*

Ključne riječi: *afazija, CroDA, AphasiaTalk, hrvatski jezik, diskurs*